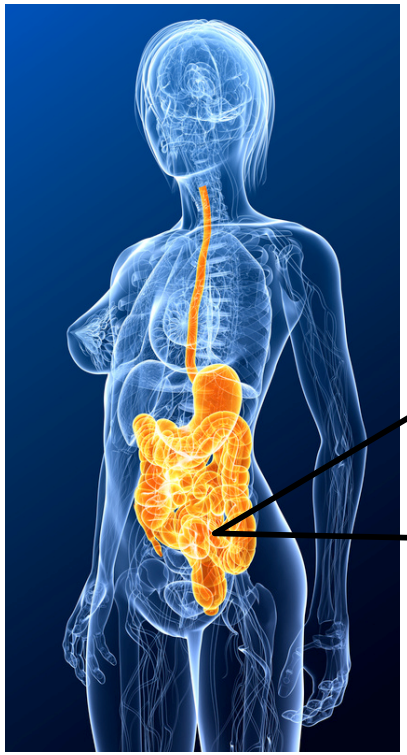
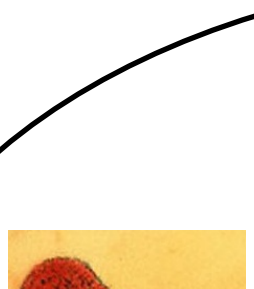
Le langage Go dans le calcul HPC Big Data

Louiza HANIS
INRA, stagiaire M2 master Haute Performance & Simulation

La métagénomique ?

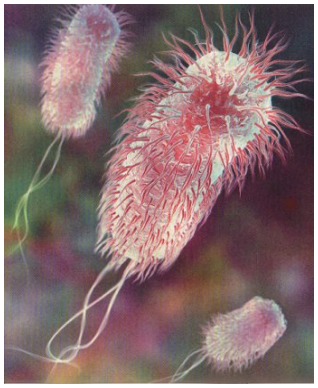




protists

archae

virus

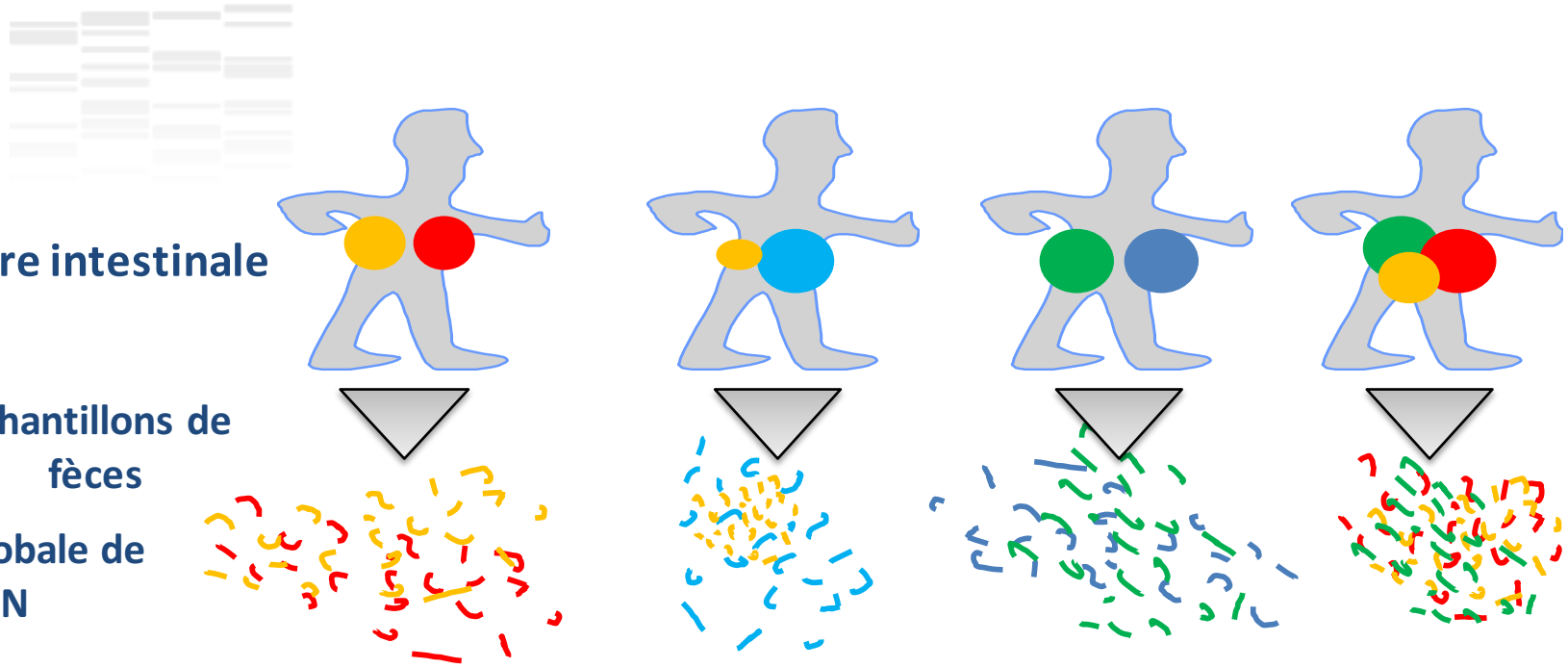


bacteria

fungus

- Big Data de l'ADN Bactérien
- 2 kg – plus de bactéries que de cellules humaines
- Comment découvrir la ou les bactéries signatures (Biomarqueurs)?

Métagénomique quantitative



Identification & Quantification

Abondance relative par projection sur un catalogue de références

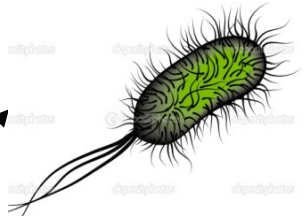
1	0	0	1
0	0	10	0
1	5	0	20
1	0	0	1
2	0	0	1
0	8	0	0
0	0	10	0
0	0	1	1
....			

Il s'agit d'extraire des clusters de manière déterministe et exhaustive

100 – 1000 individus

Aujourd'hui 10 millions de gènes

id_fragment_c	MH0001	MH0002	MH0003	MH0004	MH0005	MH0006	MH0007	MH0008
1	1.0246529439	8.3869647739	1.4874450848	2.3749503763	2.7242547919	4.7625759493	3.3844980635	1.5393655887
12	1.3467531952	1.1931356982	1.6585258148	9.8972700447	1.3267630323	1.3677934976	8.3035359532	8.5820901674
57	7.2892017460	4.3815309244	7.7707332906	2.5614971800	1.5525920094	7.7887477214	1.3471502008	1.6083954357
60	1.2526378532	1.8943414823	2.2545038192	5.3278079500	4.3180365068	5.4154756691	8.2863054205	3.5226732177
129	9.5875880325	4.5681726127	1.3502910575	2.7355209738	3.5472134610	6.7671067865	1.6093633649	4.6580802879
144	1.1328224535	5.8291787839	1.5753395670	3.6258692993	3.4426081345	3.5020169235	7.4943601770	2.4709660277
182	6.7023324214	6.7852874733	2.8881225396	1.1156520920	8.0145374559	1.1726628944	2.6077647116	2.6153179949
183	4.6178749698	0	2.2914030066	2.9504848716	0	3.1898819869	4.6551832869	1.4821164552
226	1.0106821962	4.5445646333	4.0299384274	1.2453790795	4.4732302079	7.4801496204	5.8219863330	5.5608090259
233	1.0614220079	1.0694423241	1.3794018099	4.4404063367	2.9665750583	3.4084920584	3.7364986913	7.1377548690
234	1.2627409521	9.9364209779	0	6.0509943978	0	4.9064710222	4.4552997108	0
241	1.2079685134	2.1448128208	3.6982056910	3.9183878356	4.1050069896	5.9720328692	1.0647312505	9.0761558293
305	3.8180469521	2.6821842247	2.6672661101	1.5225369727	3.3943923342	2.7671053485	2.3009688632	6.0657795514
336	4.6774055863	4.3185918062	9.5738868718	3.1558777411	7.7577180236	5.8109604316	5.3941829174	1.3078654156
378	9.0855751438	9.5325014259	0	0	0	1.1767552451	9.1589784995	4.3740510020
448	8.2490245187	4.2948632798	2.8563849293	2.0596654007	2.2898678445	8.1295252468	1.3067480313	1.9707262756
451	1.1553742081	5.1089223328	4.5303882975	4.0134307940	3.6206851565	9.2709854413	2.6261724625	1.8754101028
548	1.2977239878	8.0829117930	2.5297423705	3.8219906220	3.5568166228	1.2560645402	2.6496412253	4.3634085422
549	4.3167827556	0	3.8085132391	1.2554151014	8.4548966568	3.1811185748	5.7772018228	3.9414525513
550	6.7728832890	1.7765116293	0	0	0	1.3158263196	0	0
604	1.6843557917	2.3531803258	0	0	0	2.2734172014	6.2914030321	0
667	4.3613263603	5.9217054312	6.5639148629	2.9750722455	2.9143772567	2.7047541014	3.6982845002	4.6192629522
668	3.2878074218	6.4506344601	1.6824014794	2.0796621522	1.4939720306	1.3115679819	1.6406131390	2.6116906468
694	9.7111194218	9.5792293741	2.5483434173	2.8875701207	2.3572168888	8.5141703033	2.3930076177	1.7581969714
781	5.6827274172	0	1.1868967384	1.7116854015	1.4492040331	9.2528426127	1.1574188583	3.6849881729
793	5.8187230211	6.3653511150	8.4668087482	3.7794077387	5.0123491581	1.5715623795	2.1405755939	5.8415664848
854	6.9878954569	4.8250933143	3.0562143277	1.2592898394	1.1873388823	3.6589409122	1.0927775934	6.3257880453
888	1.1012636746	4.2537844423	2.4004184696	2.8730249126	2.5375844187	8.4336713316	2.5637285793	1.5023537973
903	6.9627772130	7.1226494299	0	0	1.7976532611	4.9599683698	2.3513751344	0
952	6.8983070536	7.2376399715	0	0	2.3153108206	1.8494669936	0	0
994	1.0974066213	4.1577931751	1.7016761281	4.0609005774	2.9592138299	6.3171149004	2.3047347697	5.8702484807
1009	7.0509488786	0	2.5408702695	3.1408387343	2.8203650871	1.2733803092	5.5061307695	0
1123	1.1108393138	4.9787587065	2.2074821959	4.6615781682	3.4304262231	9.2191652965	2.3320392708	3.9979383363
1124	9.1251421559	3.9005245355	0	1.0688882319	0	2.8890398234	1.8738429065	0
1126	8.3226108213	2.4841052444	2.2028053268	2.1934854692	1.8745867269	1.1243996092	2.3867677022	8.3588560144
1130	1.7287429374	0	0	4.4938154199	1.3450971954	0	5.2520016570	1.8811478085
1190	8.9810739477	3.8037231967	8.4324745684	3.5179686407	2.6676124671	7.0433525623	7.7661897982	1.1344862209
1191	7.7670679920	8.9640495060	1.5897922236	1.8014198918	2.6470031964	5.0902761293	1.7225601765	8.2264261965
1314	7.1794292847	9.9829516849	4.4262144401	1.4134442852	5.1587827302	1.5404522685	6.7142477490	0
1389	9.6459768114	6.2112753790	1.2710563744	2.9677982303	2.3044244616	6.7239046649	1.6985538366	8.7694910066
1390	1.6806162007	4.2671544844	2.2359658371	5.7582043463	3.3410478724	5.0282196828	1.2113487692	4.8950297814
1470	7.0550867594	2.1712919914	0	9.9169074853	1.0686049941	6.9690606030	7.3017411927	9.9631161714
1599	5.8386924905	3.3692461936	0	0	0	8.3184422503	6.4744503186	3.0920015704
1690	1.1404912664	2.2461641291	7.9672345922	4.3497608004	2.7857426743	8.7620925037	1.7696830871	1.6490675042
1798	7.2967527147	1.2955775418	1.7328735238	2.7219445517	4.8246629900	9.7826719194	1.9709514505	3.3092958244
1803	4.8897307755	2.3319365063	8.2714726673	1.1502665961	1.2394797783	3.2241408626	7.6179093248	4.2800976392
1806	2.3764502768	3.4283557760	0	0	0	3.3857519334	7.2468408829	0
1834	5.6440694075	2.6587248875	3.9294184213	6.8811194796	8.7233060745	5.2513703458	1.0643937598	0
1872	1.0741136493	7.4895225961	3.5878879127	7.8636270808	2.5420559331	8.5016678505	1.0752721454	9.4803219957
1877	8.6647871282	5.3503805265	7.9075063016	2.3505785929	2.0814809142	5.0637297088	1.9216539961	4.6762996371
1880	4.2330520556	3.9083255846	0	0	0	1.4081929443	0	0
1883	5.3574565079	2.9007103948	1.2861170684	1.7013694404	4.0573595870	2.2804229575	3.6671691258	6.8652097368
1884	3.5621072053	0	0	3.0865302836	0	1.1123219608	1.0821847794	0
1900	6.0560452767	6.2761432746	0	0	0	3.5216762811	1.7405401842	0
1942	4.7673207617	1.4229340720	0	0	9.3373251914	1.1710428410	1.3671776158	0
1958	4.3503406516	1.5259205683	0	0	0	5.3351177979	9.9318394057	4.0655537021



bactérie = ~3000 gènes

parasites = ~6000 gènes

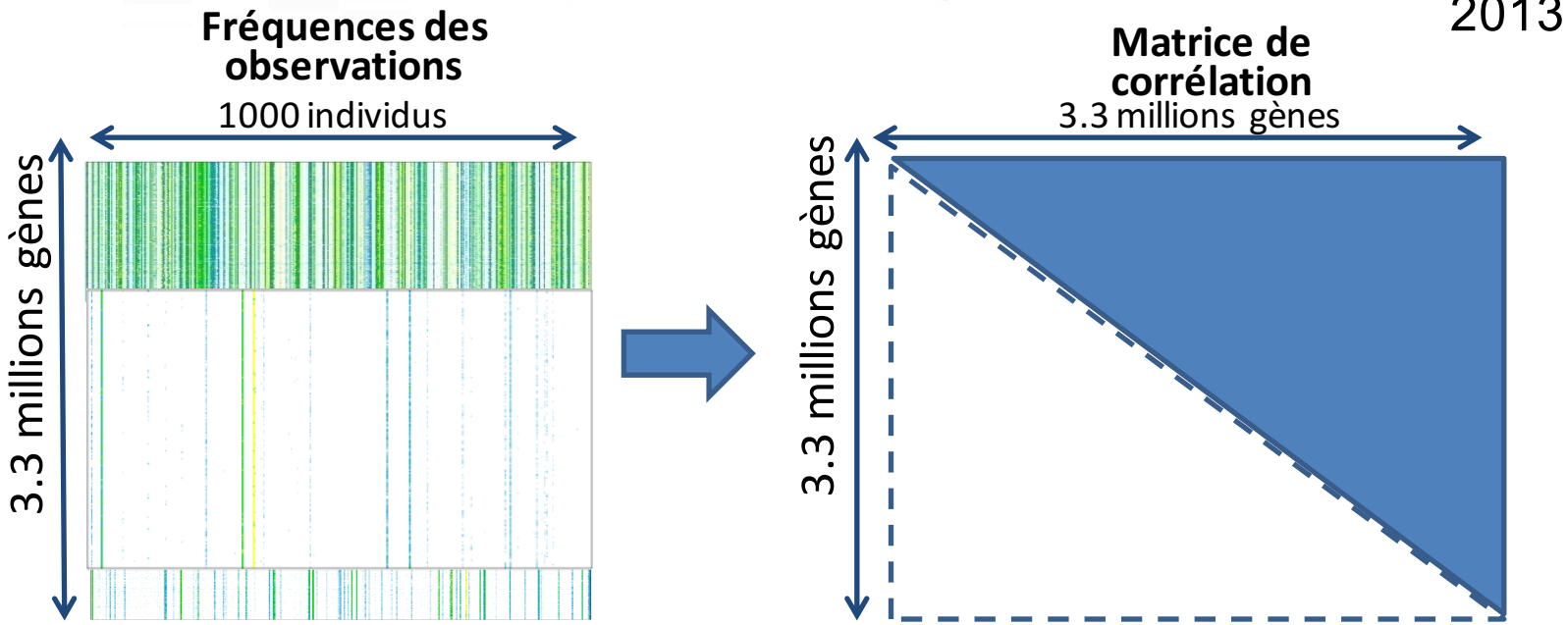
Virus = ~50 gènes

Fréquences d'observations

→métagénomique quantitative

Maintenir un code HPC spécifique?

- IN/OUT



EN 2010 un code de calcul OpenCL/CUDA a été réalisé pour traiter des matrices de taille 1000 * 3300000

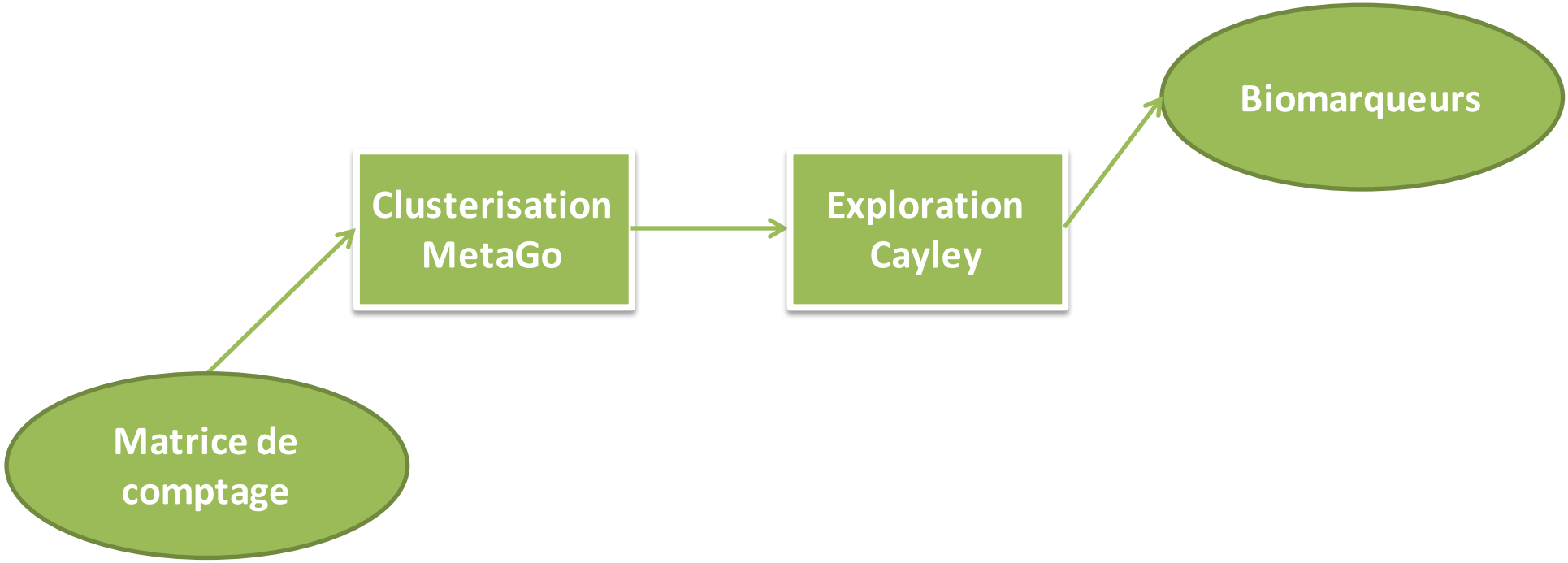
http://www.genci.fr/sites/default/files/INRA_AS_Plus_Tello_Almeida_Metaprof.pdf

→Comment maintenir du code HPC spécifique BigData (CUDA) et du code de traitement de data en amont de R?

→Peut-on unifier les codes?

Plan de développement

- ✓ Tests unitaires
- ✓ Tests d'intégration
- ✓ Tests de vérification avec R
- ✓ Benchmarks



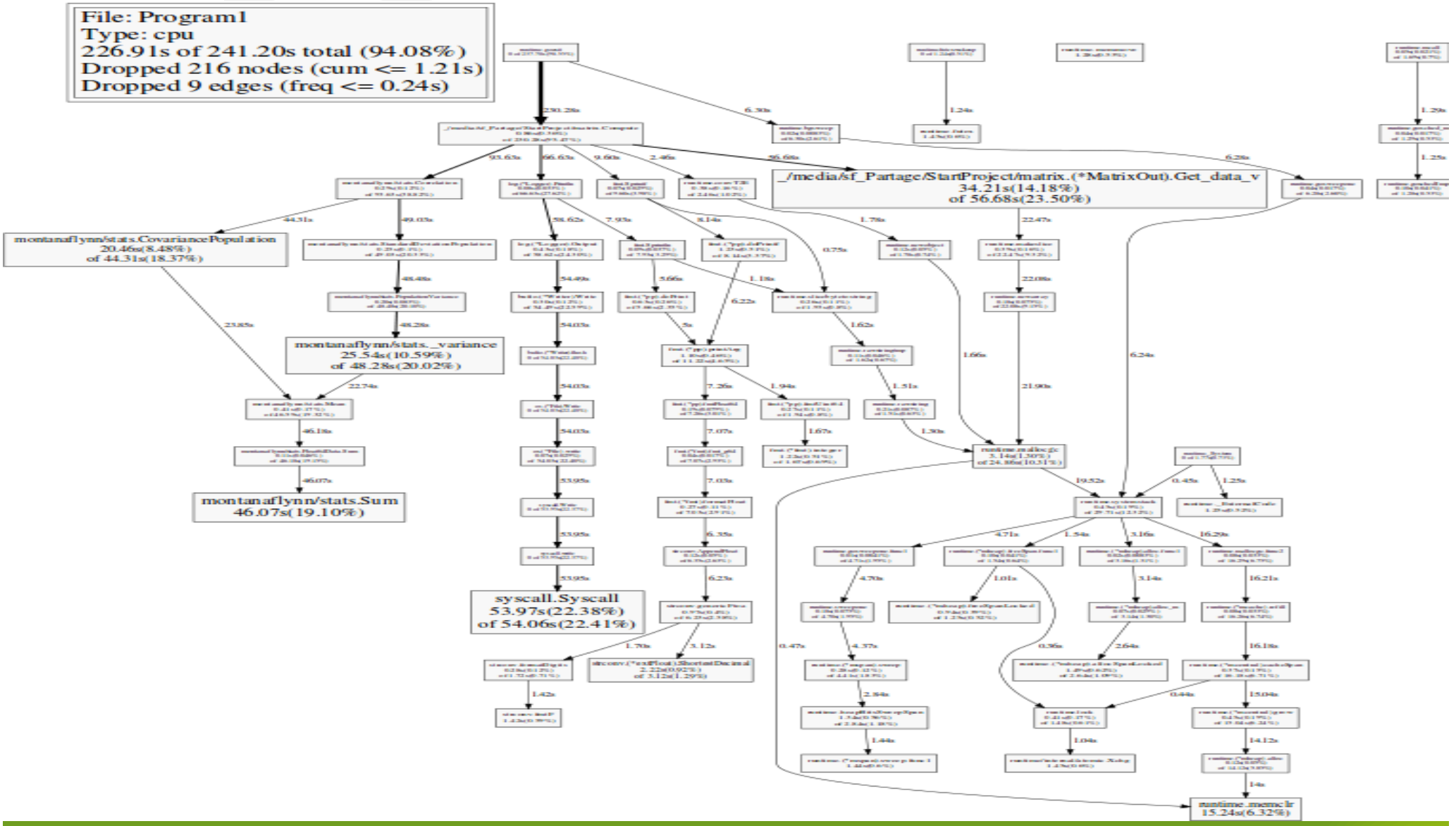
Les contraintes

Langage R

- Langage interprété ≡ Langage lent
- Pas orthogonal
- Non conçu pour le Big Data

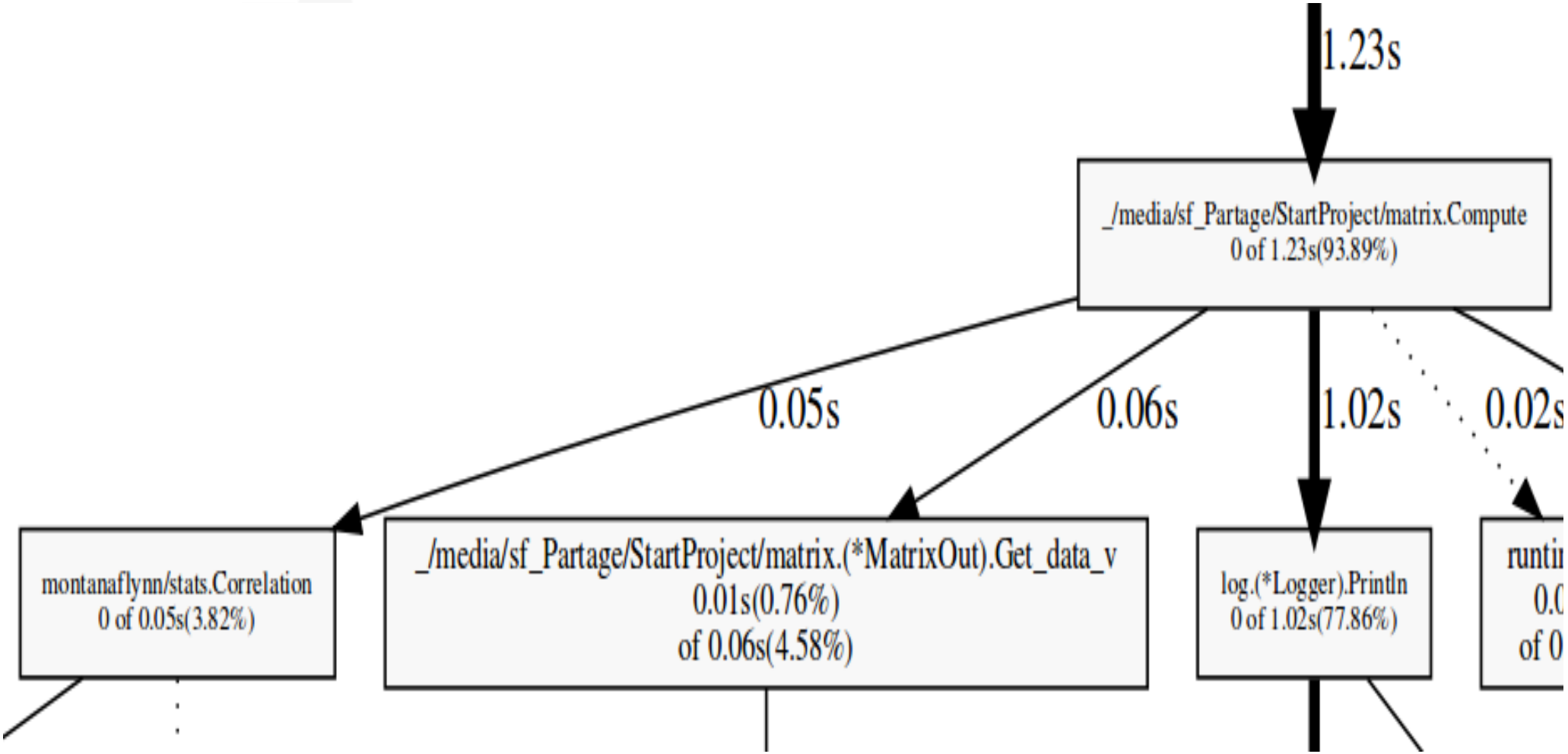
Go Profiler

- ✓ Profileur intégré



Go Profiler

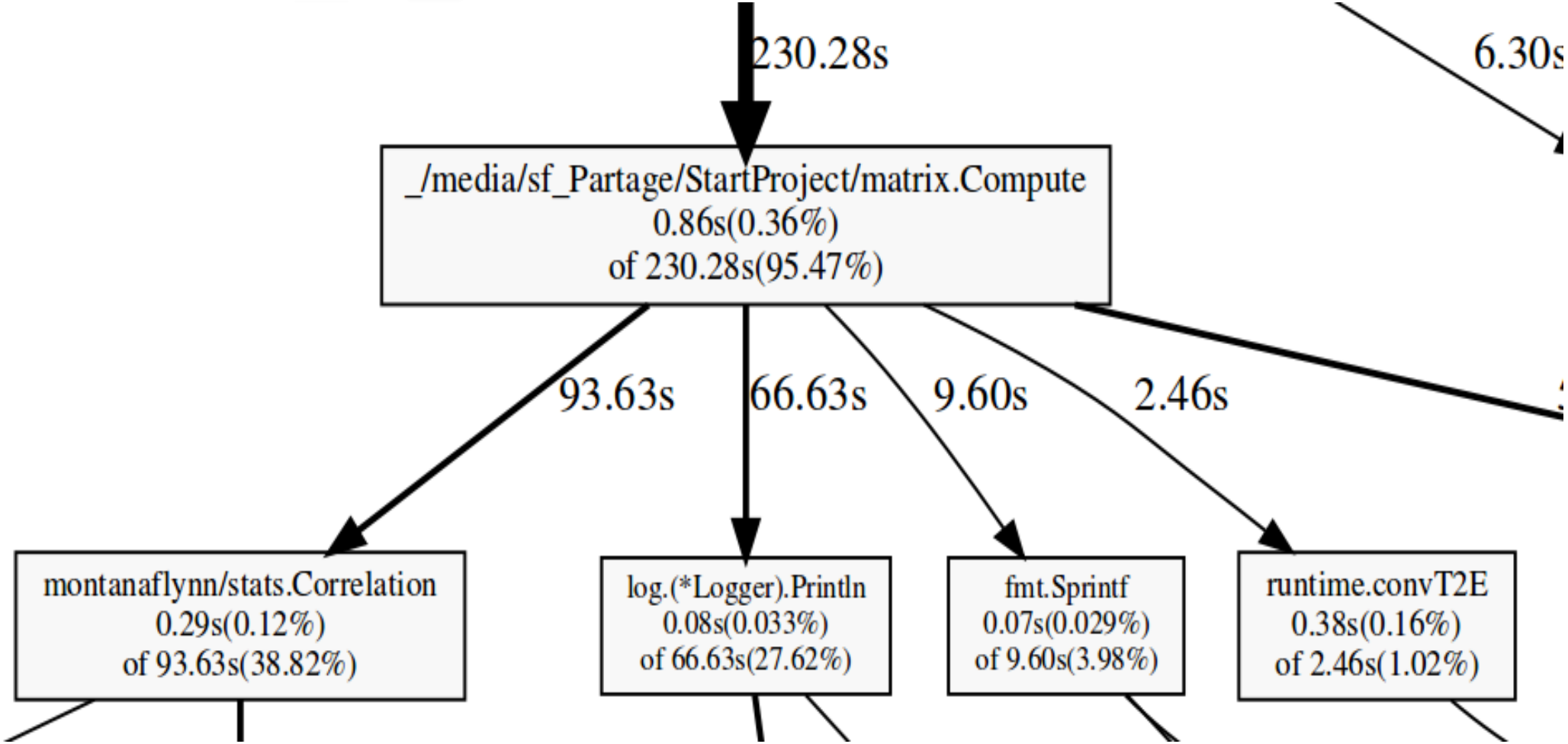
- Temps d'écriture > Temps de calcul



Matrice 1000 * 50

Go Profiler

- Temps de calcul > Temps d'écriture



Matrice 5000 * 1267

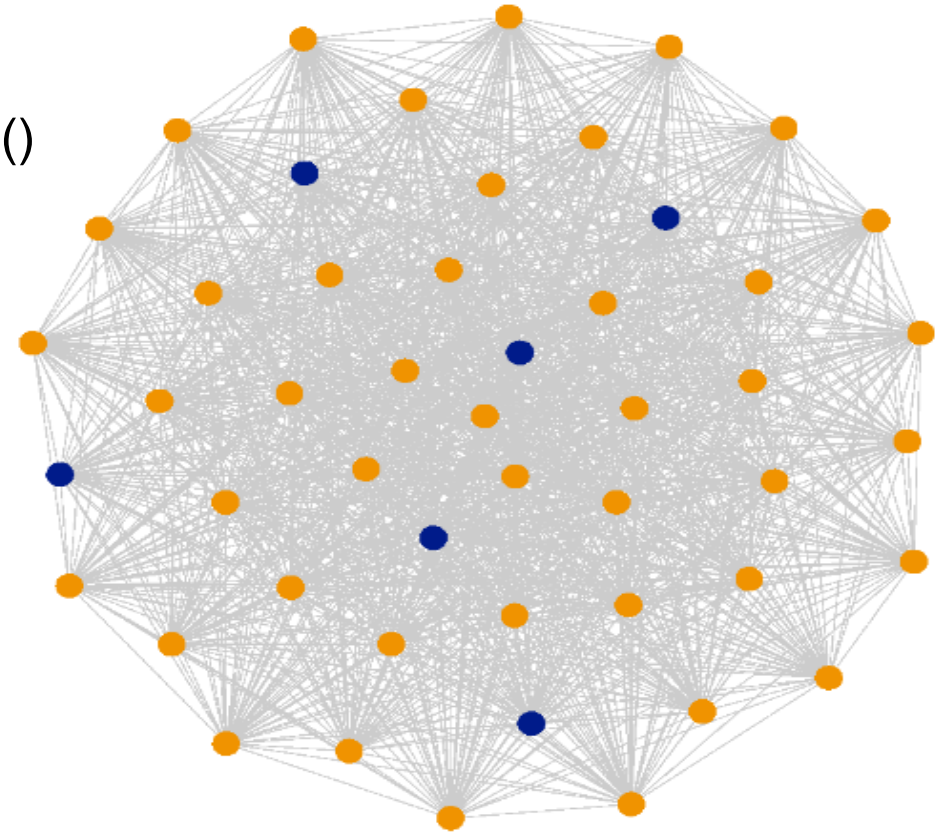
Cayley

- https://github.com/google/cayley

- Il s'agit d'un navigateur (http) et d'un moteur de requête pour une database de type "graphe orienté"
- Stockage par triplet (sujet, prédicat, objet)
  - Sujet
  - Prédicat/Objet = règle+objet (ex. Prolog)
- Un ensemble de triplet = RDF Graph (Resource Description Framework)
- Triplet+"label" = N-quads
- Repose sur Bolt (Clé/Valeur + B-tree)

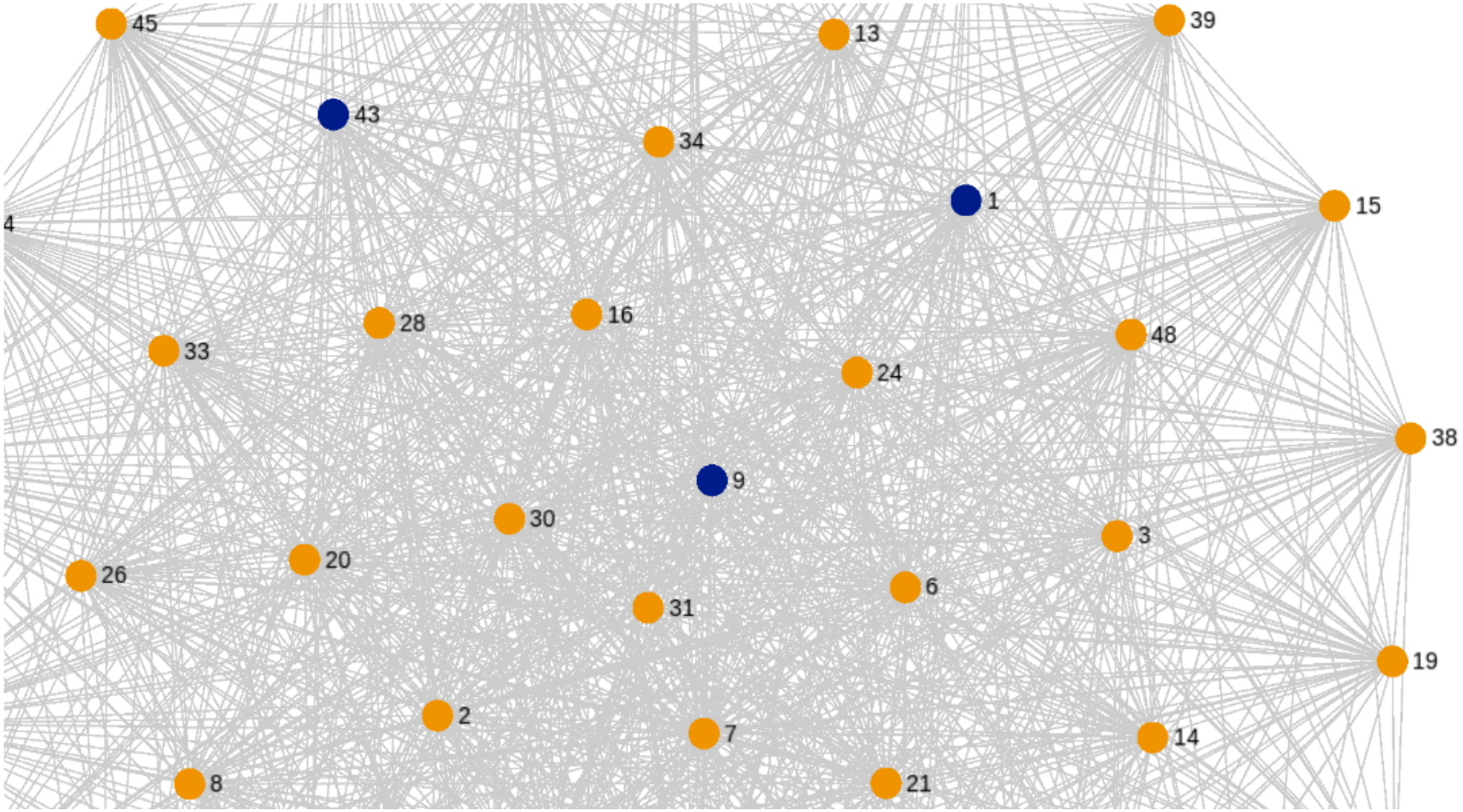
Cayley


```
g.V().Tag("source").Out().Tag("target").All()
```

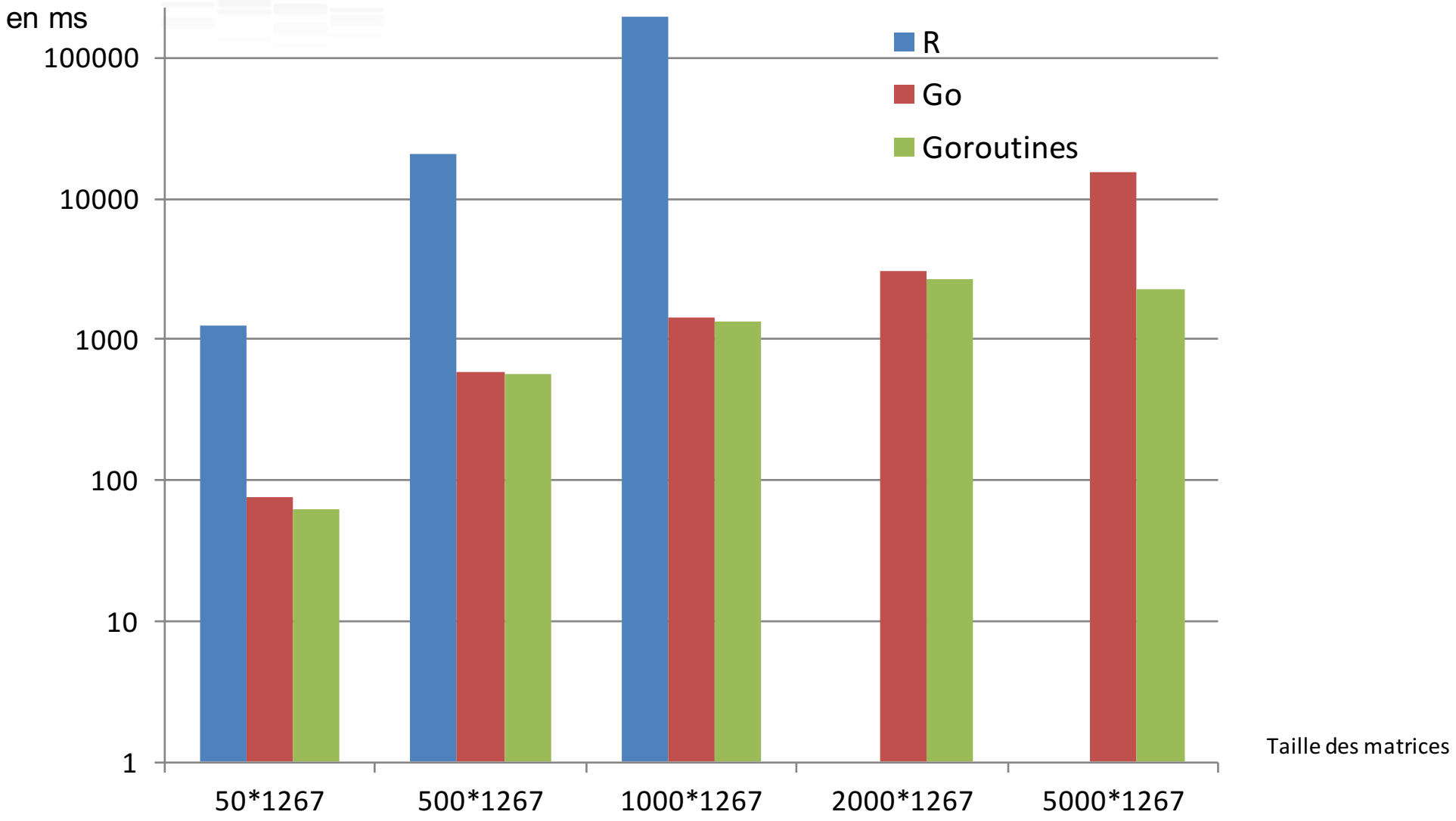


Cayley

Cayley



Le benchmark R vs Go



Avantages de développer en go

- ✓ Environnement complet
- ✓ Intégration de plusieurs outils
- ✓ Packages faciles à importer (Go get…)
- ✓ Documentation riche